Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

THE MODEL EVIDENCE DOSSIER FOR TRACEABLE, ADAPTIVE, AND CHANGE-CONTROLLED ARTIFICIAL INTELLIGENCE IN PHARMACEUTICAL DEVELOPMENT

Carlos Ramirez[1]*, Elena Torres[2], Pablo Ortega[1], Sofia Mendes[3]

1. ***Department of Model Evidence Dossier and AI Traceability, Faculty of Pharmacy, University of Barcelona, Barcelona, Spain.***
2. ***Department of Adaptive AI and Change Control, Faculty of Pharmaceutical Sciences, University of Lisbon, Lisbon, Portugal.***
3. ***Department of Evidence-Based AI Governance in Pharma, Faculty of Pharmacy, University of Porto, Porto, Portugal.***

ARTICLE INFO

Received:
21 March 2026
Received in revised form:
21 July 2026
Accepted:
23 July 2026
Available online:
28 August 2026

Keywords: Artificial intelligence governance, Pharmaceutical development, Model provenance, Change control, Lifecycle monitoring, Validation evidence

ABSTRACT

Artificial intelligence models increasingly participate in pharmaceutical activities in which data, scientific assumptions, software dependencies, intended uses, and decision consequences evolve over time. Conventional model reports are usually static: they describe a model at development or publication but rarely preserve the relationships among model version, data provenance, validation scope, uncertainty, limitations, human approvals, subsequent changes, and continuing performance. This creates an evidentiary gap when models are retrained, transferred between development stages, exposed to altered data distributions, or incorporated into decisions different from those originally evaluated. This article proposes the Model Evidence Dossier, a living, version-bound, and queryable evidence-governance architecture for artificial intelligence in pharmaceutical development. The dossier organizes model identity, intended use, data lineage and suitability, technical configuration, validation evidence, uncertainty characterization, known limitations, monitoring signals, change-impact assessments, requalification decisions, and human challenge records as connected evidence objects. Its central principle is that no isolated performance measure can establish pharmaceutical usefulness, scientific validity, continuing fitness, or acceptable reliance. Instead, each claim must remain traceable to the model version, data state, decision context, assessment method, responsible actor, and boundary conditions that support it. The architecture distinguishes monitoring from authorization, updating from requalification, confidence from calibrated uncertainty, data accessibility from suitability, and predictive performance from experimental or clinical confirmation. The proposed dossier is conceptual rather than empirically validated and does not constitute a regulatory submission, deployment protocol, or universal evidence standard. Its value will depend on future ontology development, prospective evaluation, organizational integration, and evidence that dossier use improves scientific review, change control, and pharmaceutical decision quality.

To Cite This Article: Ramirez C, Torres E, Ortega P, Mendes S. The Model Evidence Dossier for Traceable, Adaptive, and Change-Controlled Artificial Intelligence in Pharmaceutical Development. Pharmacophore. 2026;17(4):26-37. https://doi.org/10.51847/2PDeVVb1Uu

Introduction

Artificial intelligence now spans multiple drug-discovery and development tasks, but its pharmaceutical value remains conditional on well-defined questions, suitable evidence, and integration with scientific decision processes rather than algorithmic performance alone [1-3]. Models may contribute to target prioritization, molecular representation, virtual screening, translational analysis, trial planning, and quantitative pharmacology, yet these applications differ substantially in their data-generating processes, error consequences, interpretive requirements, and dependence on experimental confirmation. A favorable evaluation result in one task therefore cannot be treated as a general property of the model. It is evidence about a specific model state, dataset, endpoint, assessment design, and decision context. Pharmaceutical usefulness emerges only when those relationships remain visible to the scientists, reviewers, quality functions, and decision owners who must judge whether model outputs are appropriate for a particular stage of development.

Corresponding Author: Carlos Ramirez; Department of Model Evidence Dossier and AI Traceability, Faculty of Pharmacy, University of Barcelona, Barcelona, Spain. E-mail: carlos.ramirez@ub.edu

The difficulty is intensified by the characteristics of pharmaceutical data. Chemical structures, assay measurements, omics profiles, imaging data, electronic records, literature-derived relations, pharmacokinetic observations, and clinical endpoints differ in provenance, scale, missingness, bias, temporal stability, and biological meaning. Chemical and biological datasets may contain measurement noise, duplicated or related entities, inconsistent endpoint definitions, unrecorded preprocessing decisions, and sampling structures that complicate apparently straightforward performance claims [4]. When these conditions are incompletely documented, a model can remain computationally reproducible while becoming scientifically uninterpretable. The resulting governance problem is not simply whether an algorithm can generate an output, but whether the evidence supporting that output can be reconstructed, challenged, updated, and bounded as the model and its environment change.

Current reporting practices provide important components of transparency, but a publication, model card, validation report, dataset description, software repository, or quality record ordinarily captures only part of the required evidence. These artifacts are also frequently separated by organizational boundaries and maintained on different timescales. A model-development report may describe the training data but not subsequent drift; a validation report may present performance without preserving the complete software environment; an operational record may document an update without showing which scientific claims were invalidated; and an approval record may name the decision without retaining the objections or unresolved limitations that shaped it. Static documents consequently struggle to represent a model whose evidence state changes through retraining, data revision, workflow modification, transfer to a new population, or expansion of intended use.

This article develops the Model Evidence Dossier as an original AI evidence-governance architecture for pharmaceutical development. The dossier is proposed as a living, version-bound, relational evidence object that binds model identity, intended use, data lineage, configuration, validation, uncertainty, limitations, ownership, approvals, monitoring, change assessment, and requalification status. Its purpose is not to replace existing technical reports, quality systems, scientific judgment, or regulatory documentation. Rather, it provides an organizing architecture through which those evidence sources can be connected to precise claims and model versions. The central argument is that traceable and adaptive pharmaceutical AI cannot be governed by optimizing or reporting a single performance measure; it requires a continuing evidentiary record capable of showing what was known, for which use, about which version, under which conditions, by whom it was accepted, what subsequently changed, and whether the prior evidence remains applicable.

Why Adaptive AI Requires a Living Evidence Record

Responsible translation requires end-to-end stewardship, evidence appropriate to the local workflow and population, and explicit recognition that generalizability cannot be presumed [5-7]. Although these principles have been developed prominently in healthcare AI, the underlying problem is directly relevant to pharmaceutical science. Models used in discovery and development are embedded in evolving systems of assays, databases, laboratory practices, disease definitions, software environments, project objectives, and human decisions. A model that appears stable at the level of its stored weights may therefore undergo a meaningful evidence change when its input pipeline, compound space, assay protocol, endpoint construction, user group, or decision consequence changes. Evidence that was sufficient for an exploratory ranking exercise may be inadequate when the same output is incorporated into candidate progression, biomarker selection, clinical-trial design, or dose-related reasoning.

Continual monitoring and controlled updating are needed because the conditions under which an algorithm was evaluated may not remain stable after implementation [8]. A living evidence record must therefore do more than store the latest performance result. It should preserve the approved reference state, define the signals that may indicate deteriorating applicability, and retain the reasoning through which those signals are interpreted. Monitoring may include changes in input distributions, assay platforms, missingness patterns, subgroup composition, calibration, prediction ranges, workflow behavior, downstream confirmation rates, incidents, overrides, or user challenges. No single signal is universally diagnostic. An apparent shift may be benign, while stable aggregate performance may conceal important changes in a subgroup, chemical region, experimental process, or decision pathway. Monitoring is therefore an evidence-generation activity that requires interpretation rather than an automatic mechanism for retraining or continued authorization.

The term living should not imply that the dossier or model changes without control. In the proposed architecture, the dossier is living because it accumulates evidence while preserving prior states; it is version-bound because every claim is attached to a defined model and evidence baseline; and it is change-controlled because a substantive update cannot silently overwrite the evidence supporting the preceding version. A living record must distinguish at least four events: a change in the external environment, a change in available evidence, a change in the computational model, and a change in intended use. These events may occur independently. New validation data may strengthen or weaken confidence without altering the model. A model retraining event may alter predictions without changing the intended use. A workflow change may modify decision consequences even when both data and model remain technically unchanged. The dossier should make these distinctions explicit so that review effort can be directed toward the evidence relationships actually affected.

The living-record requirement also changes the meaning of model approval. Approval cannot be represented as a permanent label attached to an algorithmic object. It is better represented as a bounded decision that a particular evidence package supports a specified use of a specified model version under stated conditions. When those conditions change, the original approval remains historically valid but may cease to justify continuing reliance. The dossier must consequently preserve superseded versions, earlier validation packages, change rationales, unresolved limitations, dissenting assessments, and the basis for transition to a new evidence baseline. This approach does not establish which evidence threshold should apply in every

organization or pharmaceutical context. It instead proposes the informational structure needed for such thresholds and decisions to be transparent, challengeable, and proportionate. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why adaptive ai requires a living evidence record within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why adaptive AI requires a living evidence record

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Version-bound evidence	Which exact model, data state, configuration, and evaluation does a claim concern?	Model behavior and evidence applicability may change when data, parameters, software, or context change	Establishes the unit to which evidence and approval attach	A model name or release label may be treated as sufficient identification	Version identity enables traceability but does not establish validity
Intended-use continuity	Is the current decision context equivalent to the context originally evaluated?	Error consequences, users, populations, inputs, outputs, and reliance differ across uses	Determines whether existing evidence remains applicable	Evidence may be transferred to a new task or development stage without reassessment	Similar technical inputs do not establish equivalent intended use
Environmental and data change	Have data distributions, assays, workflows, populations, or endpoints changed?	Models may encounter conditions not represented in development or validation evidence	Motivates monitoring and change-impact assessment	Stable aggregate metrics may conceal local or subgroup change	Detection of change does not alone establish performance deterioration or its cause
Evidence accumulation	Has new experimental, computational, operational, or scientific evidence altered the model's support?	Evidence may strengthen, qualify, contradict, or narrow earlier claims	Makes the dossier scientifically adaptive even when the model is unchanged	New evidence may be added without reconciling contradictions	Accumulation is not equivalent to confirmation; conflicting evidence must remain visible
Controlled updating	What review is required before data, code, parameters, or intended use are changed?	Different changes affect different assumptions, validation results, and decision risks	Connects model adaptation with proportional reassessment	Retraining may be treated as routine maintenance that preserves previous validity	Technical updating does not automatically preserve scientific qualification
Monitoring evidence	Which signals indicate that evidence applicability should be reconsidered?	Performance, calibration, drift, incidents, overrides, and user feedback reveal different aspects of model behavior	Creates a continuing basis for review	Monitoring may be reduced to one metric or interpreted as automatic diagnosis	Monitoring identifies potential concerns but does not replace causal investigation
Historical evidence preservation	Can prior versions, decisions, objections, limitations, and transitions be reconstructed?	Scientific accountability requires knowledge of what was accepted and why	Prevents evidence from being overwritten by the newest model state	Superseded evidence may be deleted or detached from the version it supported	Archival completeness does not prove that historical decisions were correct
Human stewardship	Who owns monitoring, review, challenge resolution, and requalification decisions?	Evidence requires interpretation and consequences must remain attributable	Preserves accountability across organizational and model transitions	Responsibility may become diffuse or reduced to symbolic sign-off	Human oversight does not convert inadequate evidence into adequate evidence

Contents and Structure of the Model Evidence Dossier

Dataset provenance, model-development information, intended setting, input–output handling, human interaction, and failure cases should be documented as connected evidence rather than presented through an isolated endpoint measure [9-11]. The proposed dossier extends this principle from reporting into lifecycle governance. Its structure is organized around a persistent dossier identity and a sequence of immutable, version-specific evidence baselines. Each baseline contains a defined intended-use profile; a data-lineage and suitability record; a model and configuration record; a validation package; an uncertainty and calibration record; a limitation and prohibited-use register; named ownership and approval information; and links to the monitoring, incidents, challenges, and changes that follow release. The dossier is therefore neither a single report nor an

unrestricted archive. It is a relational structure in which evidence objects are connected according to the claims, versions, decisions, and dependencies they support.

Minimum-information frameworks improve transparency and critical appraisal by requiring consistent reporting of model and study characteristics [12]. However, a living dossier must represent relationships that static minimum-information checklists do not ordinarily capture. A validation result should be linked to the exact model artifact, evaluation dataset, endpoint definition, preprocessing pipeline, split strategy, software environment, assessment date, responsible analysts, and intended-use claim it supports. A limitation should be connected to the affected outputs, user groups, chemical or biological regions, decision contexts, and monitoring triggers. An approval should identify the evidence reviewed, conditions imposed, unresolved disagreements, and circumstances requiring reconsideration. These links allow the dossier to answer not only "what information exists?" but also "which claim does it support, under what conditions, for which version, and what would be affected if it changed?" **Figure 1** presents the living model evidence dossier, showing how the article's key components and boundaries are connected within contents and structure of the model evidence dossier.

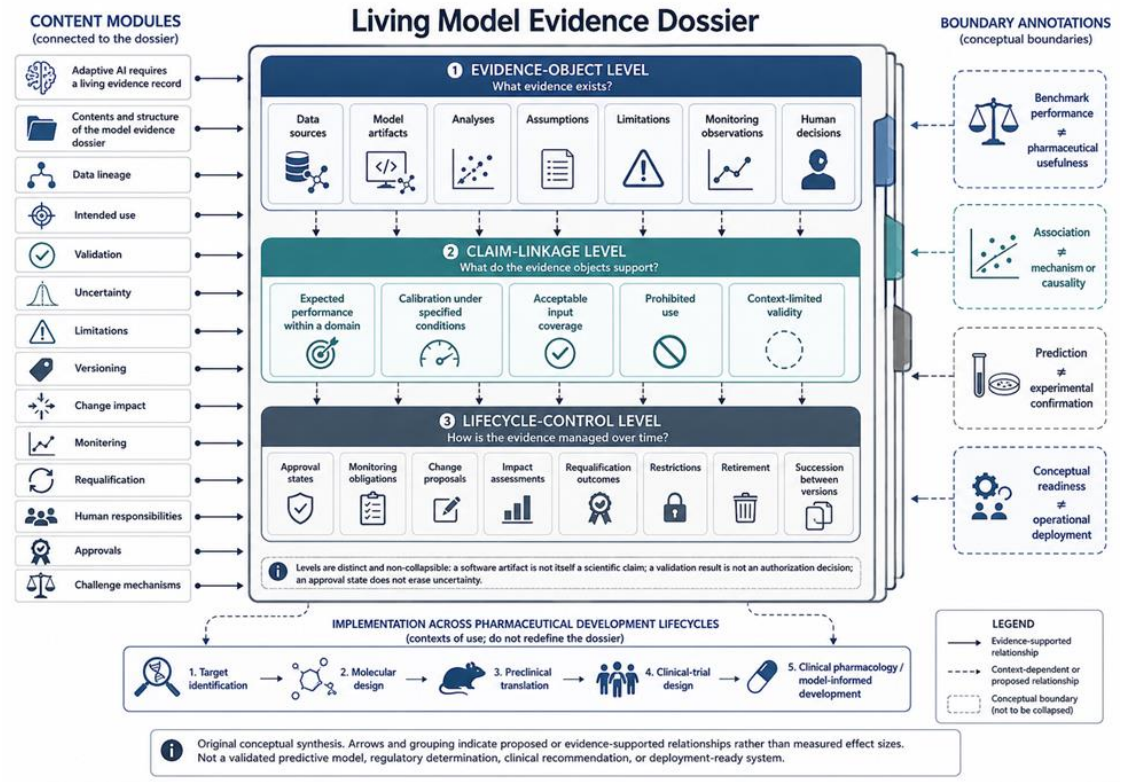

Figure 1. Living Model Evidence Dossier.

The figure is an original conceptual synthesis that organizes the article's central contribution across adaptive AI requires a living evidence record, contents and structure of the model evidence dossier, data lineage, intended use, validation, uncertainty, and limitations, data lineage, intended use, versioning, change impact, monitoring, and requalification. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

The proposed internal organization has three interacting structural levels. The first is the evidence-object level, containing identifiable data sources, model artifacts, analyses, assumptions, limitations, monitoring observations, and human decisions. The second is the claim-linkage level, which relates each object to a proposition such as expected performance within a domain, calibration under specified conditions, acceptable input coverage, or a prohibited use. The third is the lifecycle-control level, which records approval states, monitoring obligations, change proposals, impact assessments, requalification outcomes, restrictions, retirement, and succession between versions. These levels should not be collapsed. A software artifact is not itself a scientific claim; a validation result is not an authorization decision; and an approval state does not erase uncertainty. Their separation permits the dossier to preserve different forms of evidence while making their dependencies inspectable.

A practical dossier may be implemented through linked repositories, metadata services, graph-based provenance structures, quality-management records, and controlled documentation systems rather than through one software platform. The architectural requirement is logical interoperability: identifiers, version relationships, evidence dependencies, role assignments, and boundary statements must remain consistent across systems. Access should also be role-sensitive because proprietary data, personal information, confidential development strategies, and security-sensitive implementation details may not be available to every reviewer. Restricted access, however, must not produce unverifiable assertions. The dossier should indicate that protected evidence exists, who may review it, which claim it supports, and what independent assessment has been

completed. The architecture can therefore support traceability without assuming universal data openness, but its scientific value remains conditional on the accuracy, completeness, and reviewability of the records entered into it.

Data Lineage, Intended Use, Validation, Uncertainty, and Limitations

Data lineage is the foundation on which the remaining dossier claims depend. FAIR-oriented stewardship can improve the findability, accessibility, interoperability, and reuse of biomedical data, but these properties do not by themselves establish data quality, scientific appropriateness, ethical acceptability, or fitness for a particular pharmaceutical decision [13]. The dossier should therefore distinguish provenance from suitability. Provenance identifies where an observation originated, how it was collected, transformed, filtered, linked, and partitioned, whereas suitability evaluates whether those processes produced evidence appropriate for the intended question. A complete lineage record should connect raw sources to derived variables, labels, representations, model inputs, validation sets, and reported outputs. It should also preserve assay versions, inclusion and exclusion rules, endpoint definitions, normalization decisions, entity resolution, imputation, temporal cut-offs, and data-access constraints. This architecture allows reviewers to reconstruct evidence pathways without implying that complete provenance guarantees unbiased or biologically meaningful data.

Pharmaceutical data integration makes relational lineage particularly important. Drug-discovery datasets and knowledge graphs combine entities and relations from sources that vary in curation, scope, confidence, update frequency, licensing, and semantic consistency [14]. A dossier should therefore record not only which source contributed an entity or relation, but also how conflicting identities, duplicated observations, inferred links, and source-specific confidence statements were handled. The intended-use profile provides the interpretive anchor for this information. It should specify the scientific question, development stage, expected users, target population or chemical domain, allowable inputs, output meaning, decision supported, permissible degree of reliance, and prohibited extensions. Intended use is not a descriptive label added after development. It determines which data are relevant, which errors matter, what comparisons are appropriate, what validation is necessary, and which monitoring signals should trigger reconsideration.

Uncertainty must be represented as an evidence domain rather than reduced to a model confidence score. Uncertainty-quantification methods address different sources of incomplete knowledge and may behave differently across architectures, datasets, and shifts [15]. The dossier should record the type of uncertainty being estimated, the method used, the assumptions required, the evaluation design, the calibration evidence, and the limits of interpretation. Aleatoric variability, parameter uncertainty, model-form uncertainty, data sparsity, out-of-distribution inputs, assay variability, and uncertainty in the pharmaceutical decision itself should not be conflated. Even a well-calibrated predictive distribution cannot capture every unknown mechanism, omitted variable, experimental artifact, or future change. Uncertainty evidence should therefore modify how predictions are interpreted, reviewed, and escalated rather than being treated as a numerical warranty of correctness.

Validation should test claims that are explicitly linked to intended use, data lineage, uncertainty, and limitations. Leakage between model development and evaluation can inflate apparent performance and undermine reproducibility even when conventional reporting appears complete [16]. The dossier should preserve split logic, related-entity controls, temporal separation, preprocessing boundaries, tuning history, comparator selection, subgroup definitions, error analyses, and independent reproduction attempts. It should also distinguish internal evaluation, external evaluation, temporal evaluation, perturbational testing, experimental confirmation, and prospective decision utility. A limitation and prohibited-use register should identify where evidence is absent, contradictory, fragile, or inappropriate for reliance. Such a register does not neutralize the documented weakness; it makes the weakness available to users, reviewers, monitoring systems, and change-impact assessments. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop data lineage, intended use, validation, uncertainty, and limitations within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Data lineage, intended use, validation, uncertainty, and limitations

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Source provenance	Dataset identifiers, laboratory systems, databases, literature sources, acquisition dates, custodians	Establishes where evidence originated and under which collection conditions	Traceable source record linked to downstream artifacts	Original metadata may be incomplete, inconsistent, restricted, or historically unavailable	Independent source reconstruction and metadata-consistency review
Transformation lineage	Cleaning, normalization, imputation, feature construction, entity resolution, exclusions, aggregation	Shows how source observations became model-ready evidence	Ordered transformation chain with versioned parameters and responsible actors	Transformations may introduce bias or remove scientifically meaningful variation	Re-execution of transformations and sensitivity analysis
Dataset suitability	Representativeness, assay relevance, endpoint quality,	Evaluates whether available data are appropriate for the intended	Context-specific suitability statement with	Suitability is partly judgment-based and may change when	Cross-functional review against the intended-use profile

	temporal scope, missingness, coverage	pharmaceutical question	accepted and excluded uses	intended use changes	
Intended-use profile	Scientific question, users, inputs, outputs, development stage, decision consequence	Defines the scope within which evidence may be interpreted	Approved use statement, prohibited uses, and reliance conditions	Future reuse and informal extension cannot be fully anticipated	Scenario-based testing and review by scientific decision owners
Model and configuration record	Architecture, objective function, features, parameters, software, hardware, dependencies	Identifies the exact computational object to which evidence applies	Reconstructable versioned model state	Nondeterminism, inaccessible dependencies, or proprietary elements may limit reproduction	Independent computational reconstruction where feasible
Validation package	Internal, external, temporal, subgroup, robustness, calibration, and error analyses	Tests model claims against the intended use and known risks	Claim-specific evidence with acceptance conditions and unresolved findings	Results remain dependent on evaluation design, data, endpoints, and comparators	Independent review of design, leakage controls, and interpretation
Uncertainty record	Predictive distributions, calibration analyses, sensitivity tests, out-of-domain indicators	Describes quantified uncertainty and its interpretation	Decision-linked uncertainty statement and escalation rules	Quantified uncertainty cannot encompass every unknown or structural error	Calibration, interval-coverage, stress-testing, and human-interpretation assessment
Limitation and prohibited-use register	Validation failures, expert concerns, negative results, drift events, unresolved questions	Makes known weaknesses and barred uses visible and actionable	Versioned limitation statements linked to affected claims and users	Unknown failure modes remain undiscovered; documented limits may be ignored	User-comprehension testing and periodic reassessment
Evidence-linkage layer	Persistent identifiers, claim records, provenance links, reviews, approvals	Connects data, models, analyses, limitations, and decisions	Queryable evidence graph for trace reconstruction and impact analysis	Incorrect links may create an appearance of traceability without substantive accuracy	Link-integrity audits and sample claim reconstruction

Versioning, Change Impact, Monitoring, and Requalification

Algorithmovigilance extends model oversight beyond initial validation by emphasizing continuing examination of effectiveness, safety, and equity after release [17]. In pharmaceutical development, vigilance should encompass more than a deployed clinical interface. Models may influence laboratory prioritization, candidate selection, translational interpretation, trial planning, and quantitative decision-making long before direct clinical use. The dossier should therefore maintain a monitoring registry containing each monitored signal, its scientific rationale, reference state, expected range, review cadence, responsible owner, escalation condition, and evidentiary meaning. Signals may concern input composition, chemical-domain coverage, assay behavior, missingness, calibration, prediction stability, downstream experimental confirmation, subgroup error, override frequency, user-reported concerns, or changes in workflow. Monitoring evidence should remain linked to the model version and intended use under which it was generated.

Data drift illustrates why monitoring must not be reduced to aggregate predictive performance. Empirical work has shown that distributional changes may be detectable in model inputs even when commonly reported overall performance measures do not show equivalent deterioration [18]. Conversely, a detected shift may not necessarily impair the relevant decision. The dossier should therefore preserve both the signal and its interpretation. A monitoring event should initiate investigation of possible causes, affected evidence claims, subgroup or domain consequences, and the continued applicability of prior validation. It should not automatically trigger retraining, withdrawal, or continued use. The architecture separates detection, diagnosis, decision, and authorization so that an alert does not become an unreviewed model change and stable headline performance does not suppress investigation of consequential local change.

Sustainable model maintenance requires assigned ownership, continuing evidence review, and explicit decisions about whether a model should remain unchanged, be restricted, updated, replaced, or retired [19]. Every proposed change should therefore receive a versioned change record describing what changed, why it changed, who proposed it, which evidence objects may be affected, and whether the modification alters predictions, uncertainty, workflow, or intended use. Changes to data sources, labels, preprocessing, features, model architecture, parameters, software dependencies, interfaces, populations, assays, endpoints, or decision consequences may invalidate different portions of the evidence baseline. The dossier's evidence-linkage layer should support impact analysis by identifying which validation results, limitations, approvals, monitoring thresholds, and user instructions depend on the changed object.

Quality-management approaches can connect risk management, lifecycle monitoring, documentation, and controlled improvement, but their application to AI remains dependent on context and implementation quality [20]. In the proposed

architecture, requalification is a proportionate evidentiary response rather than a universal retesting package. A documentation-only correction may require no new model evaluation, whereas a changed endpoint, population, architecture, or intended use may require renewed validation and a new approved baseline. Possible outcomes include continued use without model modification, constrained reassessment, renewed validation, restriction, suspension, replacement, or retirement. Substantive approval creates a successor evidence baseline while preserving the superseded version and the rationale for transition. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop versioning, change impact, monitoring, and requalification within the article's central argument.

Table 3. Components, relationships, uncertainties, and validation needs within Versioning, change impact, monitoring, and requalification

Evaluation dimension	What must be tested	Suitable evidence or assessment	Failure signal	Interpretive limitation	Decision relevance
Version identity	Whether evidence is attached to the exact model, data, configuration, and environment evaluated	Artifact hashes, version manifests, environment records, dependency maps	Ambiguous or unreconstructable model state	Exact identity does not establish scientific validity	Determines whether earlier evidence can legitimately be reused
Monitoring coverage	Whether monitored signals address the important failure pathways of the intended use	Performance, calibration, data, subgroup, workflow, confirmation, incident, and user-feedback monitoring	Material changes occur outside monitored dimensions	No monitoring set can cover every unknown failure	Determines whether continuing evidence remains sufficiently observable
Drift sensitivity	Whether the monitoring system detects relevant distributional or contextual change	Reference-distribution comparisons, domain-coverage analysis, temporal evaluation	Delayed or missed detection of consequential change	Detected drift does not alone establish harm or model failure	Triggers investigation and possible reassessment
Alert burden	Whether monitoring produces a manageable and interpretable review load	False-alert analysis, review time, signal redundancy, escalation history	Alert fatigue or routine dismissal	Low alert volume may reflect insensitive monitoring	Affects sustainability and escalation design
Change classification	Whether a proposed change is described at the correct scientific and operational level	Dependency mapping, expert review, intended-use comparison	Substantive changes are classified as administrative or minor	Classification remains partly judgment-based	Determines the proportional reassessment pathway
Change-impact completeness	Whether all affected claims, evidence, users, and workflows are identified	Evidence-graph queries, independent review, scenario analysis	Indirect dependencies or downstream decisions are omitted	Complete dependency knowledge may be unavailable	Determines which evidence must be regenerated or qualified
Reassessment proportionality	Whether the response matches the magnitude and consequence of change	Risk-based rationale, targeted tests, renewed external or prospective validation	Cosmetic retesting after a consequential modification	No universal threshold defines sufficient reassessment	Supports decisions to continue, restrict, revalidate, or retire
Successor-baseline integrity	Whether an approved change creates a coherent new evidence state	Updated dossier modules, closed challenges, approval record, archive links	New model is released while evidence remains attached to the prior state	Formal completion does not guarantee evidentiary quality	Establishes the basis for subsequent use and monitoring
Retirement and rollback readiness	Whether use can be restricted, reversed, or transferred safely	Rollback plans, archived versions, dependency records, user communications	Inability to identify affected decisions or restore a prior state	Technical rollback may not reverse decisions already made	Supports containment when evidence applicability is lost

Human Responsibilities, Approvals, and Challenge Mechanisms

Trustworthy AI depends on interacting properties that include transparency, usability, explainability, and the capacity of stakeholders to understand and appropriately act on evidence [21]. The dossier should therefore assign responsibilities to named roles rather than to an undifferentiated organization or "human in the loop." Relevant roles may include the scientific question owner, dataset steward, model developer, independent validator, domain reviewer, quality representative, operational

owner, monitoring analyst, decision authority, and challenge adjudicator. One individual may hold more than one role in small settings, but role combinations and conflicts should be visible. Each approval should identify what was reviewed, which use was accepted, what limitations or conditions were imposed, when reconsideration is required, and who retains authority to restrict or suspend use.

Ethical concerns in health AI arise across data acquisition, model development, evaluation, implementation, access, bias, accountability, and downstream consequences [22]. A pharmaceutical dossier should not isolate ethical review in a terminal checklist. Ethical and scientific concerns may be embedded in source selection, population representation, endpoint choice, uncertainty communication, experimental burden, access restrictions, or the way model outputs redistribute attention and resources. The dossier should allow these concerns to be linked to particular evidence objects and decisions. It should also retain unresolved issues rather than forcing every review into a binary approved or rejected state. Conditional approval, restricted use, required evidence generation, and explicit dissent can provide more accurate representations of the evidentiary state.

Algorithmic participation does not remove human responsibility for decisions, especially when model outputs are embedded in professional judgment or institutional processes [23]. Approval mechanisms should therefore preserve a visible chain from evidence to decision consequence. A reviewer should be able to determine who defined the intended use, who accepted the validation strategy, who interpreted monitoring signals, who authorized a change, and who could intervene when concerns emerged. This does not imply that every scientific decision must be made manually or that human judgment is inherently reliable. Human reviewers can be biased, inconsistent, overconfident, inadequately trained, or excessively deferential to technical systems. The dossier should consequently document reviewer expertise, conflicts, rationale, disagreements, overrides, and the evidence used to resolve them.

Challenge mechanisms provide a formal route through which assumptions, data, validation, explanations, limitations, or approval decisions can be contested. Post hoc explanations may appear plausible without faithfully representing a model's internal reasoning or establishing causal validity [24]. An explanation should therefore be treated as one review artifact rather than as proof of trustworthiness. A challenge record should identify the contested claim, supporting evidence, challenger, response, adjudicator, resolution, resulting restrictions or changes, and any remaining dissent. Challenges may arise from scientists, users, data stewards, quality personnel, patients or participants where relevant, or automated monitoring systems. Their value depends on meaningful access, protection from retaliation, qualified review, and the capacity to alter decisions. A symbolic channel that records objections but cannot trigger investigation or restriction would not satisfy the proposed architecture.

Implementation Across Pharmaceutical Development Lifecycles

Clinical-trial design illustrates how AI-supported decisions may combine prediction, operational planning, participant selection, endpoint strategy, monitoring, and analysis under distinct scientific and ethical constraints [25]. A dossier instantiated for this context should identify whether the model supports feasibility assessment, enrichment, recruitment, adherence monitoring, outcome prediction, site selection, or another function. It should preserve the population and trial assumptions embedded in the data, specify how outputs enter protocol decisions, document possible exclusionary effects, and distinguish simulated or retrospective evaluation from prospective trial utility. Evidence adequate for an exploratory design comparison would not automatically support participant-level allocation or operational intervention.

Across drug development more broadly, application requirements differ according to the maturity of the scientific question and the consequences of error [26]. In target identification, the dossier may emphasize source integration, biological context, association-versus-causation boundaries, perturbational evidence, and negative findings. In molecular design, it may focus on chemical-domain coverage, assay provenance, stereochemistry, synthetic accessibility, selectivity, and uncertainty beyond the training distribution. In preclinical translation, species, dose, exposure, endpoint, and laboratory dependencies become central. In biomarker development, analytical validity, temporal behavior, disease heterogeneity, and independent replication require explicit representation. A shared dossier structure enables continuity, but the content and evidence thresholds must remain stage-specific.

Reported uses of AI in drug development are heterogeneous in task, evidence type, maturity, and public visibility [27]. Implementation should therefore proceed through bounded use cases rather than organization-wide declarations that a model is "qualified" or that an AI platform is generally ready. An initial dossier may be instantiated for one model version, one decision, and one accountable team. The implementation can then test whether users can reconstruct claims, identify limitations, detect affected evidence after a change, and resolve challenges. Expansion should depend on demonstrated interoperability, manageable maintenance burden, reliable role assignment, and evidence that the dossier improves scientific review rather than merely increasing documentation volume. Publicly described applications also cannot be assumed to represent confidential internal practices or complete evidence packages.

Clinical pharmacology and model-informed development require particular care because prediction, causal interpretation, quantitative simulation, and professional judgment may interact in high-consequence decisions [28]. A dossier in this domain should specify the pharmacokinetic, pharmacodynamic, covariate, adherence, disease-progression, or trial data used; the role of mechanistic assumptions; the intended decision; and the degree of permissible reliance. It should distinguish an AI-derived association from a physiological mechanism, a predictive interval from complete uncertainty, and retrospective performance from prospective dosing or development utility. Human oversight should involve relevant clinical, pharmacometric, statistical,

data-science, and decision expertise. **Table 4** organizes the evidence, constructs, and interpretation boundaries needed to develop implementation across pharmaceutical development lifecycles within the article's central argument.

Table 4. Evaluation, implementation, and research priorities arising from The Model Evidence Dossier for Traceable, Adaptive, and Change-Controlled Artificial Intelligence in Pharmaceutical Development

Research or implementation priority	Unresolved gap	Required methodological work	Evidence needed	Responsible actors	Expected contribution
Machine-readable dossier ontology	No shared representation connects models, data, claims, evidence, versions, decisions, and limitations across pharmaceutical systems	Ontology engineering, competency-question development, identifier design, interoperability testing	Cross-stage use cases, expert mapping, link-integrity assessment	Pharmaceutical scientists, informaticians, data stewards, quality specialists	Enables queryable traceability and impact analysis
Intended-use specification	Intended use is often broad, implicit, or detached from decision consequences	Structured use-case templates and scenario-based review	User, population, input, output, workflow, reliance, and consequence evidence	Scientific question owners, end users, validators, governance teams	Makes validation and monitoring requirements decision-specific
Data-suitability assessment	Provenance and accessibility are frequently mistaken for fitness	Domain-specific suitability frameworks and comparative expert review	Assay relevance, representativeness, endpoint quality, bias, temporal scope	Data stewards, laboratory experts, statisticians, domain scientists	Improves the evidentiary basis of training and validation datasets
Validation-profile design	Single metrics dominate heterogeneous evidence requirements	Claim-based validation profiles incorporating calibration, robustness, error, subgroup, external, and prospective assessment	Task-specific benchmarks plus realistic pharmaceutical decision evaluations	Model developers, independent validators, scientific reviewers	Aligns validation with intended use and consequence
Uncertainty representation	Confidence outputs are inconsistently calibrated and communicated	Comparative uncertainty evaluation and human-interpretation studies	Calibration, coverage, out-of-domain behavior, decision-response evidence	Data scientists, statisticians, domain experts, human-factors researchers	Supports proportionate reliance and escalation
Monitoring architecture	Relevant signals and escalation thresholds remain context-dependent	Prospective longitudinal monitoring and simulated drift studies	Input, performance, calibration, subgroup, workflow, incident, and confirmation signals	Model owners, monitoring analysts, quality functions, users	Improves recognition of changing evidence applicability
Change taxonomy and impact analysis	Organizations lack tested methods for connecting changes to affected evidence	Change classification, dependency graphs, retrospective case analysis, challenge exercises	Historical and simulated model, data, workflow, and use changes	Developers, validators, configuration managers, quality specialists	Supports proportional reassessment and prevents silent evidence reuse
Requalification methods	No universal approach determines how much evidence must be regenerated	Risk-based reassessment protocols and independent requalification studies	Updated validation, unresolved limitations, decision-consequence analysis	Independent reviewers, model owners, quality and decision authorities	Creates transparent successor evidence baselines
Human challenge and dissent	Contestability is rarely evaluated as a scientific governance function	Structured challenge simulations and organizational studies	Challenge quality, response time, resolution, decision modification, unresolved dissent	Users, reviewers, ethics experts, quality leaders, decision owners	Tests whether challenge mechanisms reveal hidden weaknesses
Cross-lifecycle handoff	Evidence fragments when models move between discovery, preclinical, clinical,	Prospective handoff studies and receiving-team reconstruction exercises	Version histories, use changes, validation transfer, missing-information records	Cross-functional pharmaceutical teams and	Preserves evidence continuity without

	and maintenance contexts			information-system owners	assuming universal qualification
Governance-effectiveness evaluation	Documentation volume is often used as a proxy for governance quality	Comparative mixed-method evaluation of dossier-supported and conventional review	Error prevention, decision quality, uncertainty recognition, review burden, timeliness	Independent researchers, organizations, scientific users	Determines whether the architecture adds practical and scientific value
Staged implementation	Full-scale adoption may create complexity before value is demonstrated	Pilot use cases, maturity criteria, burden assessment, controlled expansion	Trace reconstruction, user comprehension, maintenance effort, incident response	Program leaders, technical teams, quality functions, end users	Enables evidence-based adoption without implying operational readiness

Limitations and Boundary Conditions

The Model Evidence Dossier is a proposed conceptual architecture. It has not been prospectively evaluated, empirically compared with alternative governance approaches, or shown to improve pharmaceutical decisions. The selected literature supports many of its constituent principles, including provenance, structured reporting, contextual validation, uncertainty characterization, monitoring, maintenance, accountability, and controlled change. It does not validate their integration into the particular dossier architecture proposed here. The construct should therefore be assessed as a theory-building contribution whose coherence, usability, interoperability, maintenance burden, and scientific effects remain open questions.

The architecture also cannot define universal evidence thresholds. Pharmaceutical AI systems vary in purpose, modality, development stage, consequence, reversibility, organizational setting, and degree of human reliance. A model used to generate exploratory molecular hypotheses should not necessarily receive the same validation and approval process as one affecting clinical-trial eligibility or dose-related reasoning. Conversely, classification as exploratory should not be used to avoid documenting consequential downstream influence. Proportionality requires a transparent relationship among intended use, uncertainty, potential harm, reversibility, and evidentiary maturity. These judgments will remain partly context-dependent and cannot be fully automated by the dossier.

Traceability itself has limits. A complete chain of identifiers and documents may accurately reconstruct how a result was produced while leaving the underlying data biased, the endpoint invalid, the model scientifically inappropriate, or the human decision poorly reasoned. Restricted proprietary or personal data may limit independent scrutiny. Excessive documentation may also reduce usability, create false reassurance, or displace resources from experimental and scientific evaluation. The dossier cannot establish mechanism, causality, synthesizability, developability, safety, efficacy, clinical utility, or regulatory acceptability merely by recording a prediction or approval. Its records must be treated as evidence claims that remain open to contradiction, external validation, experimental confirmation, and revision.

Research Agenda and Implementation Priorities

The first research priority is formal representation. A machine-readable dossier ontology should define model identities, versions, data objects, transformations, intended uses, validation claims, uncertainty statements, limitations, monitoring signals, changes, approvals, and challenges. Competency questions should test whether the representation can answer practical inquiries: which evidence supports a claim, which model version produced an output, what changed between baselines, which validation results are no longer applicable, and who authorized continued use. Ontology development should involve pharmaceutical scientists, informaticians, data stewards, quality specialists, model developers, validators, and intended users so that technical interoperability does not obscure scientific meaning.

The second priority is prospective evaluation. Dossier prototypes should be tested using bounded pharmaceutical use cases and realistic change scenarios. Evaluation should examine completeness, claim traceability, reconstruction success, intended-use congruence, reviewer agreement, detection of affected evidence, challenge resolution, user comprehension, maintenance burden, and decision timeliness. Comparative studies should determine whether dossier-supported review identifies limitations, leakage, shift, uncertainty, or use mismatch more effectively than conventional documentation. No single outcome should be treated as sufficient; improved traceability may coexist with excessive workload, and faster review may coexist with weaker challenge. Evaluation should therefore combine technical, human-factors, organizational, and scientific endpoints.

The third priority is staged implementation. Organizations should begin with a limited model, decision, and evidence baseline rather than attempting immediate enterprise-wide adoption. Early implementation should establish persistent identifiers, clear intended use, minimum lineage, validation linkage, limitation records, named ownership, and a controlled change process. Monitoring and challenge functions can then be added and tested before cross-lifecycle expansion. Progression should depend on demonstrated utility, reliable maintenance, role clarity, interoperability, and the ability to restrict or retire models when evidence no longer supports use. Implementation maturity should not be inferred from the volume of completed fields or the sophistication of software; it should depend on whether the dossier supports more accurate, transparent, challengeable, and appropriately bounded pharmaceutical reasoning.

Conclusion

The Model Evidence Dossier reframes pharmaceutical AI governance from the static reporting of model performance to the continuing stewardship of version-bound evidence. Its original contribution is an architecture that connects intended use, data lineage and suitability, model configuration, validation, uncertainty, limitations, monitoring, change impact, requalification, human approval, and challenge as distinct but interdependent evidence objects. The dossier does not make an AI model trustworthy merely by documenting it, nor does it establish scientific validity, clinical utility, regulatory acceptability, or deployment readiness. Its purpose is more bounded: to make visible which evidence supports which claim, for which model version and decision context, what remains uncertain, what has changed, who accepted the resulting conditions, and when prior evidence should no longer justify continued reliance. Future work must determine whether this architecture can be represented interoperably, maintained proportionately, used effectively by pharmaceutical teams, and shown to improve scientific decision quality without creating false assurance or unproductive administrative burden.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
4. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
5. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
6. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
7. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9). doi:10.1016/S2589-7500(20)30186-2
8. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5(1):66. doi:10.1038/s41746-022-00611-y
9. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723
10. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
11. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi:10.1038/s41591-020-1034-x
12. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011-5. doi:10.1093/jamia/ocaa088
13. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: Fair enough? Eur J Hum Genet. 2018;26(7):931-6. doi:10.1038/s41431-018-0160-0
14. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac404
15. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion. 2021;76:243-97. doi:10.1016/j.inffus.2021.05.008

16. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
17. Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open. 2021;4(4). doi:10.1001/jamanetworkopen.2021.4622
18. Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun. 2024;15(1):1887. doi:10.1038/s41467-024-46142-w
19. Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools—a 360° approach to model maintenance. J Am Med Inform Assoc. 2024;31(5):1195-8. doi:10.1093/jamia/ocae036
20. Overgaard SM, Graham MG, Brereton T, Pencina MJ, Halamka JD, Vidal DE, et al. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. NPJ Digit Med. 2023;6(1):218. doi:10.1038/s41746-023-00968-8
21. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med. 2020;3(1):47. doi:10.1038/s41746-020-0254-2
22. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med. 2020;260:113172. doi:10.1016/j.socscimed.2020.113172
23. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586
24. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-6. doi:10.1126/science.abg1834
25. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40(8):577-91. doi:10.1016/j.tips.2019.05.005
26. Mak KK, Pichika MR. Artificial intelligence in drug development: Present status and future prospects. Drug Discov Today. 2019;24(3):773-80. doi:10.1016/j.drudis.2018.11.014
27. Druedahl LC, Price WN II, Minssen T, Sarpatwari A. Use of artificial intelligence in drug development. JAMA Netw Open. 2024;7(5). doi:10.1001/jamanetworkopen.2024.14139
28. Johnson M, Patel M, Phipps A, van der Schaar M, Boulton D, Gibbs M. The potential and pitfalls of artificial intelligence in clinical pharmacology. CPT Pharmacometrics Syst Pharmacol. 2023;12(3):279-84. doi:10.1002/psp4.12902